DOCUMENT RESUME

ED 384 628 TM 023 773

AUTHOR Ferrara, Steven; And Others
TITLE A Beginning Validation of Causes of Local Item Dependence in a Large Scale Hands-On Science Performance Assessment.
PUB DATE 21 Apr 95
NOTE 32p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, April 18-22, 1995).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Educational Assessment; *Prediction; Science Education; *Sciences; Test Construction; *Test Items; Test Use; *Validity
IDENTIFIERS Contextual Analysis; *Hands on Science; Local Item Dependence; Maryland School Performance Assessment Program; *Performance Based Evaluation

ABSTRACT

A study was conducted to begin a process of validating hypothesized causes of local item dependence (LID) in large-scale performance assessments. Data for the study are item level scores from 26 science tasks from the 1993 edition of the Maryland School Performance Assessment Program. Causes of high LID were hypothesized from studies by Ferrara et al. (1994) and W. M. Yen (1993) and used to predict high and low LID in multi-step items in science performance assessment tasks. The predictions were then compared to actual levels of LID in the items, as identified in correlational analyses. A summary of percentages of accurately predicted levels of LID is provided, and explanations are offered for inaccurate predictions. Prediction accuracy is offered as beginning evidence of the validity of the causes of LID hypothesized in previous studies. The accuracy with which high LID clusters can be predicted through contextual analysis also suggests the utility of contextual analysis for test developers. Nine tables present analysis data. (Contains 20 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 384 628

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

STEVEN FERRARA

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

A Beginning Validation of Causes of Local Item Dependence in a Large Scale Hands-On Science Performance Assessment

Steven Ferrara
Hillary Michaels
Maryland State Department of Education

Huynh Huynh
University of South Carolina

April 21, 1995

BEST COPY AVAILABLE

Paper presented at the annual meeting of the
National Council on Measurement in Education, San Francisco

TM023773

A Beginning Validation of Causes of Local Item Dependence in a Large Scale Hands-On Science Performance Assessment

Steven Ferrara
Huynh Huynh
Hillary Michaels

March 25, 1995

More and more often, large-scale assessments of educational achievement include short constructed response items, either as the sole vehicle for capturing student knowledge and skill, or in conjunction with multiple-choice items. Assessments that include constructed response items are generally referred to as performance assessments, even when no physically observable performance other than writing is involved. Sometimes, open-ended items are included in performance assessments as independent and discrete entities, as is the case with most typical multiple choice tests (e.g., optional performance components in commercially produced norm-referenced tests, NAEP open-ended items). In other cases, open-ended items are organized into what are often called assessment tasks (e.g., Maryland School Performance Assessment Program, the New Standards project). Assessment tasks are collections of multiple choice and/or constructed responses items which are organized around a theme (e.g., properties of substances), purpose (e.g., conducting an investigation of saline levels in water samples), or culminating activity (e.g., testing and determining ways to prevent soil erosion). Some items in assessment tasks may be developed explicitly as relatively independent and discrete entities, even though they are related through a common theme, purpose, or culmination. In an other assessment tasks, multi-step items may be closely interdependent. Some test developers argue that some level of

item interdependence is necessary to achieve assessment tasks that reflect and model effective instruction, or that are authentic (Wiggins, 1989). It is reasonable to expect levels of local item dependence (LID) among both single step items and multi-step item clusters in the same task. Further, it is reasonable to expect these levels of LID to be higher than is typical for more traditionally designed educational achievement tests.

Previous studies have documented the existence of LID in a large-scale educational performance assessment in reading and mathematics (Ferrara, Huynh, & Baghi, 1992, 1994; Yen, 1993). Other studies have documented LID in traditionally designed tests in which items were explicitly developed and selected to maximize local item independence. For example, Green and Langhorst (1986), Hanna and Oaster (1980), Nicholas and Brookshire (1987), and Sherich and Hanna (1977) found evidence of LID among reading comprehension items linked to the same reading passage. In addition, Bell, Pattison, and Withers (1988) found LID to be stronger in mathematics items than in verbal items. The earlier studies focused primarily on identifying LID in tests where LID was not expected to exist. The more recent studies focus on the degree of LID (i.e., levels of inter-item correlations controlled for ability levels) and hypothesized causes of LID.

Both classical true score theory and item response theory usually require that items in a test satisfy the condition of local independence. The notion of local or conditional item independence concerns the statistical independence between examinee responses to two items for

examinees at a single point on the underlying ability scale (see Crocker & Algina, 1986, pp. 342 ff.; Lord & Novick, 1968, pp. 398 ff., 538 ff.). Examinees of equivalent ability may be able to respond successfully to a locally independent item whether or not they have responded successfully to other items in the test or even encountered them. Conversely, the ability of examinees of equivalent ability to respond successfully to a locally dependent item is correlated with -- or dependent on -- their ability to respond successfully to one or more other items. Locally independent items are necessary in order to provide scores that differentiate students on different test objectives and that are reliable and generalizable beyond the context of the test. In addition, LID may decrease the accuracy of latent trait model item and person ability parameter estimates (Ackerman, 1987; Ackerman & Spray, 1986). Often, independent testlets are formed from dependent items (e.g., Wainer & Kiely, 1987; cf. Haladyna, 1992; Huynh, 1994). However, testlets result in the loss of information about student ability and complicate item banking (Yen, 1993) and test scoring. Consequently, a goal for developers of large-scale performance assessments is to balance the need to allow some interdependence among items in a performance assessment task with the need to minimize LID. To accomplish this goal, test developers need to be aware of both hypothesized and validated causes of LID. The purpose of this study is to begin a process of validating hypothesized causes of LID.

The Maryland School Performance Assessment Program

The data for this study are item level scores from 26 science tasks from the 1993 edition of the Maryland School Performance Assessment Program (MSPAP). MSPAP is part of a larger

school reform effort. Performance on MSPAP is reported annually in "school report cards" along with information from other tests and about attendance, promotion, and graduation rates. Schools are expected to meet standards for satisfactory and excellent performance. School performance standards and other major features of MSPAP (e.g., the assessed learning outcomes, assessment activities, scoring criteria) are intended to represent high but attainable standards. The learning outcomes that form the basis for MSPAP are considered to represent what students should know and be able to do by the year 2000, not their current levels of knowledge and skill. The lowest performing schools that fail to make progress towards these standards or that decline in performance are required to undergo a self-initiated reconstitution process following guidelines and requirements established by the Maryland State Board of Education.

Approximately 55,000 students at each of grades 3, 5, and 8 participate in the administration of MSPAP each May. MSPAP contains assessment tasks and items in reading, writing, language usage, mathematics, science, and social studies and is administered for nine hours over five consecutive days. All assessment activities in MSPAP require students to construct short to medium length responses in the form of lists, phrases, sentences, paragraphs, one extended essay, sketches, diagrams, and tables. Student responses are scored using activity specific keys, abbreviated generic rubrics called "rules," or rubrics. MSPAP task developers are trained to understand item dependence, the importance item independence, and to restrict inter-dependence to the steps within multi-step items as much as possible. Further information about MSPAP is available from the Maryland State Department of Education.

Method and Procedures

The process for validating hypothesized causes of LID is based on contextual analysis procedures and the average within cluster correlation method described in other studies (Ferrara, Huynh, & Baghi, 1994; Huynh & Michaels, 1995; Huynh, Michaels, & Ferrara, 1995). In these studies, item clusters displaying high and low LID were identified using inter-item correlations within equal ability examinee groups. Then contextual characteristics (i.e., content and response requirements) were identified in an attempt to differentiate locally independent and dependent item clusters. Finally, hypothesized explanations were made for high levels of LID where it existed. The present study uses hypothesized causes of high LID from Ferrara et al. (1994) and Yen (1993) to predict high and low LID in multi-step items in science performance assessment tasks. These predictions are then compared to actual levels of LID in the items, as identified in correlational analyses. A summary of percentages of accurately predicted levels of LID is provided and explanations for inaccurate predictions are offered. Prediction accuracy is offered as beginning evidence of the validity of the causes of LID hypothesized in previous studies.

Data Base

The 1993 edition of MSPAP includes three non-parallel test forms at each of grades 3, 5, and 8. Each test form contains three self-contained science assessment tasks (two tasks in one grade 3 form) with both single step and multi-step items. Table 1 displays information on total numbers of items per form and numbers of single step items, multi-step item clusters, and

numbers of items per cluster for each form. The contextual and correlational analyses in this study are based on the 62 multi-step item clusters in 26 science assessment tasks (see Table 1).

Insert Table 1 about here

Nine data sets, one for each test form for each grade, were systematically extracted from the 1993 statewide MSPAP data file to serve as the data based for this study. Students are randomly assigned to test forms for MSPAP test administrations; approximately 6,000 cases per test form (roughly one third of all cases with complete science data) were randomly selected for analysis. Table 2 contains raw score descriptive statistics from the 1993 administration of MSPAP. The data indicate that the science assessments are difficult for examinees of average ability. The average percent of maximum score in the upper panel of Table 2 indicates that, on average, students achieved one quarter to one third of the maximum possible points available in the science assessments in each of the grade 3, 5, and 8 test forms. In addition, student raw scores are positively skewed. The lower panel of Table 2 contains raw score ranges for examinee groups of approximately equivalent ability. Raw scores of zero were deleted in the grouping of examinees to avoid inflating correlations for the lowest ability group (Huynh et al., 1995). In addition, about 3% of the cases were deleted from top score groups because they were widely scattered in the tails of the positively skewed raw score distributions. Most score groups were formed to have a raw score range of three units. In a few situations an additional score was included in a group to increase the group size to at least 150 examinees, a group size necessary to minimize sampling error. Score groups number from 8 to 10 and include approximately 95% of all examinees selected for the study.

Insert Table 2 about here

Hypothesized Causes of LID

An examination of previous work on LID resulted in identification of 11 logically derived hypothesized causes of LID (Yen, 1993) and four empirically derived hypothesized causes (Ferrara et al., 1994). After combining causes identified in both studies and eliminating causes that are not test based (i.e., instructional, test administration, and scoring causes hypothesized by Yen), four hypothesized causes of LID were listed on a recording sheet and given coding numbers for use in contextual analyses. These hypothesized causes appear in multi-step items that require examinees to compare/contrast, answer/explain (i.e., describe the process to arrive at a response, defend the answer given), and use given information. They are listed in Table 3. These four causes were derived during studies of the reading and mathematics performance assessments (Ferrara et al., 1994; Yen, 1993) and from previous work on reading comprehension items linked to the same passage (e.g., Green & Langhorst, 1986; see above) and on math and verbal items (e.g., Bell et al., 1988). The other 11 hypothesized causes listed in Table 3 were generated in the course of conducting the current contextual analyses. They appear to be related to the nature of the hands-on science performance assessment tasks and, therefore, were not identified in previous studies of LID.

Insert Table 3 about here

Procedures for Contextual Analyses and for Predicting LID

The context -- that is, content and responses requirements -- of multi-step items in each MSPAP science assessment task was analyzed by the first author. For each step in a multi-step item the question was posed, "Could examinees respond successfully at this step whether or not they were able to respond successfully to, or had even attempted, other steps in the same item?" This question was answered with consideration for the content knowledge, skills, and response format required at an item step. Each time the answer was "yes" a code corresponding to a hypothesized cause of LID from Table 3 was entered onto a recording sheet. If more than one cause was apparent, multiple codes were recorded. Item clusters with at least one item step considered to be dependent on another step within the same multi-step item are predicted to be locally dependent item clusters.

Procedures for Identifying Levels of LID

The procedures followed in this study for determining levels of LID in multi-step item clusters closely follow those described in Ferrara et al. (1994). In order to determine levels of LID, it was necessary to first establish that responses in a multi-step item are independent of responses in other multi-step items. This between cluster independence was verified by examining between cluster correlations. Then, levels of dependence within each multi-step item cluster in all tasks were established using average within cluster item correlations. Finally, levels

of LID (high, low, neither) were identified using average within cluster item correlations. Further details on procedures for identifying levels of LID are available in Ferrara et al. (1994).

Procedures for Evaluating Evidence for Hypothesized Causes of LID

The process of beginning validation of currently hypothesized causes of LID includes comparisons of predicted levels of LID and empirically identified levels of LID for each cluster and examining frequencies with which hypothesized causes were coded in the contextual analyses. First, multi-step item clusters were identified as either high or low based on contextual analyses. Next, they were identified as either, high, low, or neither based on the correlational analyses. Then, agreement between the levels of LID identified by both procedures was indicated and tabulated. In addition, the frequency with which hypothesized cause of LID was coded during contextual analysis was tabulated twice. First, code frequencies were tabulated to summarize the results of contextual analyses. Second, code frequencies were tabulated only for those clusters in which the contextual analysis accurately predicted level of LID, using the LID levels identified in the correlational analyses as the criterion for prediction accuracy.

Results

Between Cluster Correlation Analyses

Table 4 contains between cluster item correlations across all multi-step items in all tasks, pooled across the 8 to 10 score groups described previously. Most correlations in this table are

small in absolute value and negative. This result is expected because the examinee groupings are based on responses to all items in the assessment. In addition, the correlational method used in this study is similar in theoretical underpinnings to the Q3 statistic and produces highly similar results (Huynh & Michaels, 1995; Huynh, Michaels, & Ferrara, 1995), which has been shown to be negatively biased (Yen, 1993). It may be noted that most correlations are $\leq$.03 (97% of 1,757 correlations) and few exceed .11. These correlations suggest that the clusters are locally independent. They also suggest that . 03 can be used as the upper limit for low LID clusters and that .11 can be used as the lower limits for high LID clusters. These criterion values are similar to those used in Ferrara et al.(1994). They are applied to average within cluster correlations below as a means for identifying high and low LID clusters.

Insert Table 4 about here

Within Cluster Correlation Analyses

Table 5 contains within cluster item correlations pooled across the 8 to 10 score groups described previously. These correlations are strikingly different from the between cluster correlations. Only 25% (141 of 546) of these correlations are below .03 and 40% (226 of 564) exceed .11. These results suggest the existence of high LID in large numbers of item clusters with high LID.

Insert Table 5 about here

Predictions and Statistical Identification of High and Low LID Item Clusters

Tables 6-8 display results from the contextual analysis of item clusters that provided information for predicting high and low LID clusters. The two columns under "Predicted LID" contain the numbers of LID causes coded for each item cluster during contextual analysis and an indication of the predicted level of LID. Item clusters with at least one cause code are predicted to be high LID clusters. Only five of the 62 item clusters are predicted to be low LID clusters; all others are predicted to be high LID clusters. The numbers of coded LID causes increase across grades: 28 coded causes in the grade 3 assessments, 46 at grade 5, and 92 at grade 8. Item clusters at grade 3 tend to be associated with 1-3 LID causes; clusters at grade 5 with 1-3 causes, but with 1-2 clusters per form associated with 4-5 causes; clusters at grade 8 tend to be associated with 3-4 causes, but with some clusters with 6-7 causes. Numbers of causes coded also vary by form: 8, 14, and 6 causes in the three grade 3 forms; 20, 13, and 13 in the grade 5 forms; and 43, 20, and 29 in the grade 8 forms.

Insert Tables 6-8 about here

Tables 6-8 also display average within cluster (AWC) correlations, or statistically identified LID. The two columns under "Identified LID" contain AWC correlations and an indication of the identified level of LID. The criteria applied to the AWC correlations allows for identification of three levels of LID: low (AWC $\leq$ .03), medium (AWC $>.03 \leq .11$), and high (AWC $>$ .11). Only eight of the 62 item clusters across all three grades are identified as low LID

clusters; 28 are identified as high LID clusters. The numbers of low and high LID clusters are similar across grades: 2 low and 8 high at grade 3, 4 low and 9 high at grade 5, and 2 low and 11 high at grade 8. All AWC correlations except one are positive. The highest AWC correlation is .484 at grade 3 (form 3C, item cluster 3), .341 at grade 5 (form 5C, cluster 4), and .530 at grade 8 (form 8A, cluster 8).

The last column in Tables 6-8 indicates whether predicted and identified levels of LID are in agreement. Contextual analyses allow for two predicted levels of LID (high and low); the correlational analyses allow for three identified levels (low, medium, and high). Agreement between predicted and identified LID is calculated here for identified low and high LID levels; identified medium LID levels are treated separately. Level of LID was accurately predicted for 9 of 10 grade 3 clusters with high or low identified LID, for a 90% accuracy rate. For the 10 grade 3 clusters with identified medium LID, high LID was predicted for eight clusters, low LID was predicted for two clusters. At grade 5, level of LID was accurately predicted for 9 of 13 clusters with high or low identified LID, for a 69% accuracy rate. For the six grade 5 clusters with identified medium LID, high LID was predicted for five clusters, low LID was predicted for one cluster. At grade 8, level of LID was accurately predicted for 12 of 13 clusters with high or low identified LID, for a 92% accuracy rate. For the 10 grade 8 clusters with identified medium LID, high LID was predicted for all 10 clusters. Low LID was predicted for three grade 3 clusters, one grade 5 cluster, and one grade 8 cluster; low LID was identified in one each of the grade 3 and 5 clusters, while medium LID was identified for the other of these three clusters. Overall, contextual analyses accurately predicted LID levels 83% (30 of 36) of the time in 62 item

clusters. In the other 26 clusters with identified medium LID, contextual analyses predicted high LID 23 times and low LID three times.

Frequencies of Causes of LID

The frequency with which each LID cause is coded in contextual analysis reflects the design of the MSPAP science assessments but also provides beginning evidence of the validity of these hypothesized causes. Table 9 contains frequencies by and across grades for each LID cause code. As can be seen in Table 9 under the heading "All coded causes," items that require students to use given and generated information, special knowledge, and learning from the assessment task are coded for LID most often (68 times), followed by the four variations of answer/explain items (42 times). In addition, examinees are often asked to draw conclusions from an experiment (coded 15 times), replicate their responses (coded 11 times), describe patterns in data or what happens in an experiment (coded 10 times), compare and contrast (coded 9 times), and extend their thinking beyond data to give a response (coded 8 times). Roughly speaking, examinees are asked to compare and contrast and answer and explain their answers with equal frequency at grades 3, 5, and 8; to describe patterns in data, describe what happens in an experiment, and replicate responses with equal frequency at grades 5 and 8, but not at all at grade 3; to use given and other information more often at grade 8 than at grades 3 and 5; and to draw conclusions from an experiment and extend beyond data frequently at grade 8, but not at all at grades 3 and 5.

Insert Table 9 about here

Table 9 also displays frequencies of LID causes in item clusters for which high LID was accurately predicted in contextual analysis. As before, items that require examinees to answer and explain their answers and use given and other information are coded most often (19 and 21 times each). Items that require an answer and then a defense of the answer do not dominate the variations on answer/explain items as before. In contrast, items that require examinees to use information that they generate clearly dominate the other variations on this type of item in high LID clusters. Items that require examinees to describe patterns in data and to reiterate/re-explain previous answers do not appear in the accurately predicted high LID clusters; all other LID causes appear 1-3 times in these clusters. Roughly speaking, LID causes appear in these clusters with equal frequency across grades 3, 5, and 8 except for items that require an answer with a generic explanation, which appears most frequently at grade 5.

Discussion and Conclusions

This study was undertaken to begin the process of validating hypothesized causes of LID in assessment tasks in a large scale performance assessment. Hypothesized causes of LID were drawn from previous studies and derived from contextual analyses of tasks in a large scale hands-on science performance assessment. Large numbers of hypothesized causes of LID were

identified in these assessment tasks. The contextual analyses accurately predicted high LID, which was identified in correlational analyses, with 83% accuracy (30 of 36 high LID clusters) across grades 3, 5, and 8. Contextual analyses also over-predicted high LID in (a) 26 medium LID multi-step item clusters used in this study (42% of 62 total clusters), and (b) 6 low LID clusters (17% of the high LID clusters, 10% of all clusters).

The accuracy with which high LID clusters can be predicted through contextual analysis suggests at least two conclusions. First, that contextual analysis appears to be useful for test developers, as part of the task development, review, and revisions process, for minimizing LID. In fact, the contextual analyses used in this study tend to over-predict high statistical LID. From a test developer's point of view, this conservatism is preferable in that it should lead to revision of item clusters to achieve local independence, even when may not be likely to be identified in statistical analysis. Second, and most important for this study, prediction accuracy provides evidence to support the validity of some hypothesized causes of LID. Specifically, Table 9 suggests somewhat strongly that items that require students to answer and explain, describe what happens in an experiment, or use information, especially given and generated information, do elicit examinee responses that are locally dependent. This beginning validity evidence appears most strongly for causes of LID hypothesized in previous studies (i.e., Ferrara et al, 1994; Yen, 1993; add RC cites). Other, somewhat weaker evidence of the validity of previously unidentified causes of LID also appears, specifically, for items that require examinees to draw conclusions from an experiment, extend beyond available data to respond, or replicate responses. A previous

study (Ferrara et al., 1994) suggested that compare/contrast items should cause LID. The evidence in this study to support the validity of this hypothesis is relatively weak.

The number of high LID clusters in MSPAP science tasks is not surprising, given the MSPAP task design features of coherence and relatedness to a theme or purpose. One surprise in these results is the frequency with which apparently locally dependent item clusters are not identified as high LID clusters in statistical analysis. Of the 62 total multi-step item clusters, 28 (45%) were identified as high LID clusters, but 26 (42%) were identified as medium LID clusters and 8 (13%) were identified as low LID clusters. The amount of LID may seem low given the multiple LID causes coded for many clusters. This result may be due to a lack of sensitivity in correlational approaches to identifying LID in general or specifically in performance assessments. On the other hand, this results may indicate that apparent, judgmentally identified LID does not always result in dependence among examinee responses. Further, it seems plausible that some examinees may be able to formulate responses in apparent dependent situations by guessing at or approximating a reasonable response, even if they were unable to formulate an adequate response in the item step on which the response appears to be dependent. In fact, evolving conventional wisdom in scoring performance assessments probably decreases LID in seemingly dependent item clusters. For example, some MSPAP answer/explain items are scored as single units rather than as two separate and dependent items, thus creating mini testlets.

The frequency with which LID cause codes were used in this study is a function of the design of data source, the Maryland School Performance Assessment Program (MSPAP).

Conclusions from this study about the likelihood with which item cluster types appear to reflect dependency actually show LID in statistical analysis must be tentative. Several replications, including studies in which numbers of item cluster types are balanced, would provide additional evidence to support the validity of the LID causes identified in this study. However, results from this study are likely to be generalizable to tests of other content areas, grades, and item formats, including multiple choice items. Several of the hypothesized causes of LID have appeared in this study of a science performance assessments, studies of reading and mathematics large scale performance assessments (e.g., Ferrara et al., 1994; Michaels, 1995), and earlier studies of reading and mathematics multiple choice tests (e.g., Green & Langhorst, 1986).

References

Ackerman, T. A. (1987). The robustness of LOGIST and BILOG IRT estimation programs to violations of local independence. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Ackerman, T. A., & Spray, J. A. (1986). A general model for item dependence. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Bell, R. C., Pattison, P. E., & Wither, G. P. (1988). Conditional independence in a clustered item test. Applied Psychological Measurement, 12 (1), 15-26.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rhinehart and Winston.

Ferrara, S., Huynh, H., & Baghi, H. (1992). Assessing local dependency in examinations with clustered free-response items. Paper presented at the annual meeting of the National Council on Measurement in Education.

Ferrara, S., Huynh, H., & Baghi, H. (1994). Contextual characteristics of locally dependent open-ended item clusters on a large-scale performance assessment. Manuscript submitted for publication.

Green, D. R., & Langhorst, B. H. (1986). Passage dependence and item characteristics. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Haladyna, T. M. (1992). Context-dependent item sets. Educational Measurement: Issues and Practice, 11 (1), 21-25.

Hanna, G. S., & Oaster, T. R. (1980). Studies of the seriousness of three threats to passage dependence. Educational and Psychological Measurement, 44, 583-596.

Huynh, H. (1994). On equivalence between a partial credit item and a set of independent Rasch binary items. Psychometrika, 59, 111-119.

Huynh, H., & Michaels, H. (1995, April). Statistical procedures to identify local item dependency. In H. Huynh (Chair), Technical advances in partial credit models and their applications to performance assessments. Invited symposium conducted at the annual meeting of the National Council on Measurement in Education, San Francisco.

Huynh, H., Michaels, H., & Ferrara, S. (1995). A comparison of three procedures to identify item clusters with local dependency. Manuscript under preparation.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Michaels, H., Ferrara, S, and Huynh, H. (1995, April). A beginning validation of causes of local item dependence in a large scale mathematics performance assessment. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Nicholas, L. E., & Brookshire, R. H. (1987). Error analysis and passage dependence of test items from a standardized test of multiple-sentence reading comprehension for aphasic and non-brain damaged adults. Journal of Speech and Hearing Disorders, 52, 358-366.

Scherich, H. H., & Hanna, G. S. (1977). Passage-dependence in the selection of reading comprehension test items. Educational and Psychological Measurement, 37, 991-997.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computer adaptive testing: A case for testlets. Journal of Educational Measurement, 24 (3), 185-201.

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. Phi Delta Kappan, 71, 703-713.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. Applied Psychological Measurement, 8, 125-145.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. Journal of Educational Measurement, 30 (3), 187-213.

lidscims

Table 1

Numbers of Items, Item Clusters, and Tasks in the Nine Forms of the 1993 MSPAP Science Assessments

Test form	3A	3B	3C	5A	5B	5C	8A	8B	8C
Numbers of:									
Tasks	2	3	3	3	3	3	3	3	3
Items	25	27	23	28	22	24	32	27	25
Single step items	6	5	8	7	9	8	5	7	8
Multi-step item clusters	7	8	5	7	6	6	10	7	6
Numbers of steps in each item cluster									
Cluster 1	3	3	6	4	3	2	2	4	2
Cluster 2	3	3	2	2	2	2	3	3	6
Cluster 3	4	3	3	3	2	4	4	2	2
Cluster 4	2	3	2	2	2	3	3	2	2
Cluster 5	3	2	2	6	2	2	3	4	2
Cluster 6	2	2		2	2	3	2	2	3
Cluster 7	2	3		2			3	3	
Cluster 8		3					2		
Cluster 9							2		
Cluster 10							3		

Table 2

Raw Score Descriptive Statistics for Ability/Score Groups for the 1993 MSPAP Science Assessments

Test form		3A	3B	3C	5A	5B	5C	8A	8B	8C
All examinees										
	N	6116	5713	5934	5856	5606	5615	4965	4899	4637
	MPS	37	39	35	48	36	42	47	46	39
	Mn	10.83	10.82	9.25	15.96	13.26	14.67	17.63	13.22	13.56
	SD	6.15	6.75	6.02	7.40	6.63	7.79	9.34	8.19	6.64
	% of MPS	29.27	27.75	26.44	33.25	36.82	34.93	37.52	28.75	34.76
	Skewness	0.27	0.62	0.62	0.27	0.23	0.26	0.30	0.41	0.25
Ability/score group										
1	Range	1-3	1-3	1-3	1-4	1-3	1-3	1-4	1-3	1-3
	n	613	728	972	225	299	305	267	449	169
2	Range	4-6	4-6	4-6	5-7	4-6	4-6	5-7	4-6	4-6
	n	888	1022	1185	492	658	541	385	591	497
3	Range	7-9	7-9	7-9	8-10	7-9	7-9	8-10	7-9	7-9
	n	1038	921	1083	724	796	732	557	640	679
4	Range	10-12	10-12	10-12	11-13	10-12	10-12	11-13	10-12	10-12
	n	1034	912	876	874	871	772	567	647	772
5	Range	13-15	13-15	13-15	14-16	13-15	13-15	14-16	13-15	13-15
	n	916	687	685	819	888	734	569	585	720
6	Range	16-18	16-18	16-18	17-19	16-18	16-18	17-19	16-18	16-18
	n	710	540	468	818	760	654	525	557	618
7	Range	19-21	19-21	19-21	20-22	19-21	19-21	20-22	19-21	19-21
	n	450	391	310	679	611	612	528	409	507
8	Range	22-25	22-25	22-25	23-25	22-24	22-24	23-25	22-24	22-24
	n	250	294	167	530	391	511	442	343	347
9	Range				26-28	25-28	25-27	26-28	25-28	25-28
	n				313	241	370	356	328	212
10	Range				29-32		28-31	29-33	29-32	
	n				236		230	414	152	

Note. MPS = maximum possible score; Mn = average raw score; % of MPS = Mn/MPS; Range = ability/score group's raw score range

Table 3

Logically and Empirically Derived Hypothetical Causes of LID

Hypothesized cause	Source
Responding to a locally dependent item requires the examinee to:	
Compare/contrast	Ferrara et al., 1994
Answer/explain:	
- Describe process to get answer	- Ferrara et al., 1994; Yen, 1993
- Defend answer	- Current study
- Generic	- Current study
- Why something happened	- Current study
Describe:	
- Patterns in data	- Current study
- What happens in an experiment	- Current study
Use:	
- Given information	- Ferrara et al., 1994; Yen, 1993; see also previous studies cited in text
- Generated/collected information	- Current study
- Special knowledge	- Current study
- Learning from task	- Current study
Reiterate-re-explain/summarize previous steps or responses	Current study
Draw conclusions from an experiment or from given or collected data	Current study
Extend beyond data or given information	Current study
Replicate responses	Current study

Table 4

Frequency Distributions for Between Cluster Correlations Pooled Across 8-10 Ability/Score Groups for the 1993 MSPAP Science Assessments

Correlation	3A	3B	3C	5A	5B	5C	8A	8B	8C
-0.24 to -0.21			1						
-0.20 to -0.17			4				2		
-0.16 to -0.13	3	2	5	4		4	7	6	2
-0.12 to -0.09	11	26	15	17	16	13	32	17	15
-0.08 to -0.05	53	71	21	61	51	42	117	37	49
-0.04 to -0.01	71	80	16	91	49	59	178	88	48
0.00 to 0.03	30	38	18	32	19	23	81	55	20
0.04 to 0.07		2		5		3	18	7	1
0.08 to 0.11		4				1	10		
0.12 to 0.15		1					5		
Number of correlations	168	224	80	210	135	145	450	210	135

Note. Total 1,757 correlations.

Table 5

Frequency Distributions for Within Cluster Correlations Pooled Across 8-10 Ability/Score Groups for the 1993 MSPAP Science Assessments

Correlation	3A	3B	3C	5A	5B	5C	8A	8B	8C
-0.12 to -0.09				1				1	
-0.08 to -0.05	1			2	2				
-0.04 to -0.01	1	1		4	3	2	2	5	4
0.00 to 0.03	22	9	5	8	12	10	18	22	6
0.04 to 0.07	16	16	4	19	7	10	16	19	11
0.08 to 0.11	9	11	5	7	14	9	11	8	5
0.12 to 0.15	2	6	3	16	1	8	10	6	3
0.16 to 0.19		11	1	3	3	10	15		8
0.20 to 0.23		2	3	2	10	1	6		5
0.24 to 0.27		3	1		2	1	8		2
0.28 to 0.31	5		4	4		2	2		1
0.32 to 0.35		2	2	2			2		2
0.36 to 0.39		3	1	1					2
0.40 to 0.43			1			4	1	2	
0.44 to 0.47			4			1	2	1	1
0.48 to 0.51			1			1		1	1
0.52 to 0.55			1				1	3	1
0.56 to 0.59			2				2	1	
0.60 to 0.63			2				1	1	
0.64 to 0.67							2		
0.68 to 0.71									
0.72 to 0.75							1		
Number of correlations	56	64	40	69	54	59	100	70	52

Note. Total 564 correlations.

Table 6

Predicted and Identified LID Levels for the 1993 MSPAP Science Assessments, Grade 3

Grade/ form	Item cluster	N of items	Predicted LID N	Predicted LID Level	Identified LID AWC	Identified LID Level	Agreement?
3A	1	3	1	High	0.052	---	---
	2	3	1	High	0.087	---	---
	3	4	2	High	0.054	---	---
	4	2	2	High	0.222	High	Yes
	5	3	1	High	0.014	Low	No
	6	2	0	Low	0.017	Low	Yes
	7	2	1	High	0.031	---	---
3B	1	3	1	High	0.187	High	Yes
	2	3	3	High	0.062	---	---
	3	3	3	High	0.124	High	Yes
	4	3	2	High	0.060	---	---
	5	2	2	High	0.277	High	Yes
	6	2	1	High	0.124	High	Yes
	7	3	2	High	0.099	---	---
	8	3	0	Low	0.067	---	---
3C	1	6	3	High	0.084	---	---
	2	2	0	Low	0.069	---	---
	3	3	1	High	0.484	High	Yes
	4	2	1	High	0.351	High	Yes
	5	2	1	High	0.303	High	Yes

Note. N = number of LID cause codes. AWC = average within-cluster correlation. Predicted LID level: 1 or more LID cause codes. Identified LID level: AWC $\leq$ .03 is low, AWC > .11 is high. Total 28 causes coded.

Table 7

Predicted and Identified LID Levels for the 1993 Science Assessments, Grade 5

Grade/ form	Item cluster	N of items	Predicted LID N	Predicted LID Level	Identified LID AWC	Identified LID Level	Agreement?
5A	1	4	5	High	0.095	---	---
	2	2	2	High	-0.027	Low	No
	3	3	3	High	0.098	---	---
	4	2	2	High	0.276	High	Yes
	5	6	4	High	0.063	---	---
	6	2	2	High	0.115	High	Yes
	7	2	2	High	0.097	---	---
5B	1	3	1	High	0.023	Low	No
	2	2	5	High	0.067	---	---
	3	2	1	High	0.180	High	Yes
	4	2	2	High	0.140	High	Yes
	5	2	2	High	0.151	High	Yes
	6	2	2	High	0.020	Low	No
5C	1	2	1	High	0.126	High	Yes
	2	2	0	Low	0.067	---	---
	3	4	3	High	0.114	High	Yes
	4	3	3	High	0.341	High	Yes
	5	2	2	High	0.178	High	Yes
	6	3	4	High	0.015	Low	No

Note. N = number of LID cause codes. AWC = average within-cluster correlation. Predicted LID level: 1 or more LID cause codes. Identified LID level: AWC $\leq$ .03 is low, AWC > .11 is high. Total 46 LID causes coded.

Table 8

Predicted and Identified LID Levels for the 1993 MSPAP Science Assessments, Grade 8

Grade/ form	Item cluster	N of items	Predicted LID N	Predicted LID Level	Identified LID AWC	Identified LID Level	Agreement?
8A	1	2	3	High	0.224	High	Yes
	2	3	4	High	0.050	---	---
	3	4	7	High	0.025	Low	No
	4	3	6	High	0.112	High	Yes
	5	3	6	High	0.052	---	---
	6	2	3	High	0.182	High	Yes
	7	3	4	High	0.127	High	Yes
	8	2	3	High	0.530	High	Yes
	9	2	4	High	0.089	---	---
	10	3	3	High	0.244	High	Yes
8B	1	4	5	High	0.044	---	---
	2	3	0	Low	0.020	Low	Yes
	3	2	3	High	0.057	---	---
	4	2	3	High	0.478	High	Yes
	5	4	4	High	0.078	---	---
	6	2	3	High	0.039	---	---
	7	3	2	High	0.031	---	---
8C	1	2	5	High	0.043	---	---
	2	6	5	High	0.045	---	---
	3	2	3	High	0.157	High	Yes
	4	2	5	High	0.184	High	Yes
	5	2	4	High	0.341	High	Yes
	6	3	7	High	0.134	High	Yes

Note. N = number of LID cause codes. AWC = average within-cluster correlation. Predicted LID level: 1 or more LID cause codes. Identified LID level: AWC $\leq$ .03 is low, AWC > .11 is high. Total 92 LID causes coded.

Table 9

Distribution of LID Cause Codes Showing Most Frequently Occurring Likely Causes of LID

LID cause	Frequencies: All coded causes 3	5	8	Total	When LID accurately predicted 3	5	8	Total
Compare/contrast	3	3	3	9	0	1	1	2
Answer/explain:								
- Describe process	2	1	6	9	0	1	3	4
- Defend answer	6	8	10	24	3	3	3	9
- Generic	2	6	0	8	1	4	0	5
- Why something happened	1	0	0	1	1	0	0	1
Describe:								
- Patterns in data	0	3	4	7	0	0	0	0
- What happens in an experiment	0	3	0	3	0	2	1	3
Use:								
- Given information	2	5	11	18	1	1	2	4
- Generated information	9	7	12	28	5	3	5	13
- Special knowledge	2	2	9	13	1	1	0	2
- Learning from task	0	3	6	9	0	1	1	2
Reiterate/re-explain	1	0	2	3	0	0	0	0
Draw conclusions from an experiment	0	0	15	15	0	0	2	2
Extend beyond data	0	0	8	8	0	0	3	3
Replicate responses	0	5	6	11	0	1	1	2
Total	28	46	92	166	12	18	22	52

\steve\aera95\lidsci tables